Unsupervised Learning Based on Multivariate Libby-Novick Beta Mixture Model for Medical Data Analysis

Niloufar Samiee

A Thesis
In the Department of
Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Applied Science (Information Systems Security)
at Concordia University
Montréal, Québec, Canada

February 2024

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Niloufar Samiee

Entitled: Unsupervised Learning Based on Multivariate Libby-Novick Beta Mixture Model for Medical Data Analysis

and submitted in partial fulfillment of the requirements for the degree of

"Master of Applied Science (Information System Security)"

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Farnoosh Naderkhani — Examiner and Chair

Dr. Roch Glitho — Examiner

Dr. Nizar Bouguila — Thesis Supervisor

Approved by _____
Dr. Jun Yan (GPD) — Chair of Department or Graduate Program Director

Dr. Mourad Debbabi — Dean of Faculty of Engineering and Computer Science

Date March 20, 2024

Abstract

Unsupervised Learning Based on Multivariate Libby-Novick Beta Mixture Model for Medical Data Analysis

Niloufar Samiee

This thesis proposes a set of innovative clustering techniques that leverage finite and infinite mixture models to analyze medical data and images of cells. The proposed approaches are designed to improve the accuracy and efficiency of clustering in these domains. These models utilize a flexible distribution, the Libby-Novick Beta distribution, to better model data with varying shapes due to an additional shape parameter compared to the conventional Beta distribution. In this study, our initial approach involves the use of deterministic learning techniques, with a focus on maximum likelihood using the expectation-maximization approach. To achieve accurate data representation in unsupervised learning, it is crucial to determine the optimal number of clusters. So, we expand the minimum message length (MML) principle to ascertain the number of clusters in Libby-Novick Beta mixtures. In order to overcome the challenge of estimating the number of mixture components, we extend our finite mixture model to an infinite one. Nonparametric Bayesian techniques can effectively capture data distribution with an unknown number of components. This approach is useful for complex data sets and can lead to more accurate predictions and better decision-making. Our models are evaluated for different medical applications throughout the entire process, and they consistently show superior performance over traditional alternatives. This study reveals the significance of the Libby-Novick Beta distribution and the recommended mixture models in converting medical data into practical insights. This conversion aids healthcare professionals in making more accurate decisions, thereby advancing the overall healthcare field.

Acknowledgements

I would like to convey my sincere appreciation to Prof. Nizar Bouguila, my supervisor, for his constant assistance, direction, and motivation throughout my graduate studies. The privilege of being supervised by him has proven to be exceptionally invaluable, and the substantial knowledge I have acquired from his expertise and sagacity cannot be overstated.

Additionally, I would like to extend my heartfelt appreciation to Dr. Narges Manouchehri, for her continuous support, both academically and emotionally. Her insights, advice, and understanding have been instrumental in my progress, and I am grateful for her assistance and the positive impact she has had on my academic career.

Finally, I would like to extend my deepest thanks to my family and friends for the persistent encouragement and support they have provided throughout this challenging yet rewarding academic endeavor. Their unwavering belief in my potential has been a profound source of strength and motivation. I am immensely grateful for the pivotal role they have fulfilled in my educational journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and Related Work

In today's world, data play an increasingly critical role, which makes the need to analyze various types of data inevitable. Machine learning is becoming more prevalent in analyzing complex data in a wide range of fields as a result [1–7]. The annotation of large datasets is a time-consuming and expensive process in many fields, including medical datasets. In such circumstances, unsupervised methods are very useful. A number of sophisticated unsupervised machine learning techniques are available to analyze data, including Principal Component Analysis (PCA), Anomaly Detection (AD), Autoencoders, and clustering. The clustering method is one way of dividing data points into groups based on similarities between them. The increasing dimension and sparsity of data sets make clustering more challenging. Finite mixture models are highly regarded statistical learning techniques for clustering [8–13]. These models allow to determine the probability of each data point to be assigned to a given cluster or component, making them particularly suitable for modeling complex datasets. To effectively describe the model's components, it is critical to identify the most appropriate probability distribution. Gaussian mixture models (GMM) are frequently utilized for clustering tasks due to their exceptional flexibility in modeling complex data distributions in diverse and challenging conditions [14, 15]. However, in practical situations, the assumption of a Gaussian distribution may not be suitable, particularly in cases where the data are asymmetric and do not conform to the Gaussian distribution [16–21]. Therefore, it is important to

identify alternative probability distributions that can effectively model the data to obtain accurate and meaningful results. It has been demonstrated that alternative distributions, such as Beta [19, 22], Dirichlet [23–27], and McDonald's Beta [28], may provide superior performance than Gaussian distributions for data clustering in a wide range of applications [28].

The Libby-Novick beta distribution is employed in the development of a new finite mixture model proposed in this thesis. Our reason for choosing this distribution is based on its increased flexibility in comparison to the conventional Beta distribution, due to an additional shape parameter. It is noteworthy to highlight that the Libby-Novick distribution, as a member of the Beta distribution family, includes the Gaussian distribution as a special case when specific parameters are chosen. Accordingly, the Libby-Novick distribution reduces to the Gaussian distribution under certain circumstances, demonstrating its flexibility and inclusivity in accommodating a broad range of data patterns. This allows us to achieve a good fit even for non-Gaussian and asymmetric data, with highly promising results observed in real-world datasets. As such, our approach represents a compelling alternative to the traditional Gaussian distribution.

When using mixture models, we confront two main challenges. The first challenge involves learning model's parameters. In addition, estimating the complexity of the model, which takes into account the number of clusters, is the second problem. Model's parameters can be estimated using a variety of methods including deterministic and Bayesian approaches. Deterministic approaches, such as maximum likelihood (ML) estimation via the expectation-maximization (EM) algorithm are among the most commonly used techniques for estimating model's parameters [29] due to their simplicity and low computational complexity. The EM algorithm, despite its advantages, is known to have problems with convergence to local maxima and dependence on initialization [30]. Furthermore, these deterministic techniques have limited capacity to assess the model's complexity. As a result of these limitations, clustering tasks may not be as effective as they should be.

By contrast, Bayesian methods are more flexible and do not suffer from the limitations of deterministic methods. In the Bayesian framework, Bayes' theorem is employed to determine probability distribution properties from data [31]. We combine our prior beliefs about parameters with knowledge

derived from observations in order to determine posterior probability. Sampling is a key component of Bayesian approaches. Monte Carlo Markov Chain (MCMC) sampling algorithms can be used for Bayesian inference to generate samples from a probability distribution [32, 33].

Despite the fact that Bayesian learning of finite mixtures overcomes a great deal of the limitations associated with deterministic methods, we still face the challenge of selecting the optimal number of clusters. Indeed, it is essential to determine the optimal number of clusters in a data set in clustering, and this is the case in many fields, including medical applications, particularly when it comes to making sense of large, complex datasets frequently encountered in healthcare and medical research. In that context, we propose an infinite multivariate Libby-Novick Beta mixture model based on a Bayesian framework, using a mixture of Dirichlet processes to extend the finite mixture model to infinity. The use of nonparametric Bayesian methodologies allows for an infinite increase in the number of mixture components. Thus, it overcomes the model's selection problem and makes the resulting model more flexible and practical for use in real-world problems.

1.2 Contributions

The main contributions of this thesis are as follows:

- We introduce a new finite mixture model based on the Libby-Novick Beta distribution, which is an extension of the Beta distribution. This approach provides a flexible and superior means of fitting non-Gaussian and asymmetric data. The proposed approach addresses the limitations of traditional mixture models and offers an improved alternative for modeling complex data. We propose a finite mixture model with a focus on deterministic learning techniques to estimate the parameters of finite mixture models such as ML via EM using Newton Raphson's method.

- We tackle a crucial issue in unsupervised learning: identifying the ideal number of clusters to accurately represent the data. Our approach involves expanding upon the minimum message length (MML) princi-

ple, allowing us to determine the optimal number of clusters in Libby-Novick Beta mixtures.

- We propose an extension of the finite multivariate Beta mixture model to the infinite case. Our proposed model is developed within a nonparametric Bayesian framework that uses the Markov Chain Monte Carlo technique to estimate the posterior distribution. This approach allows for flexible and scalable modeling of the underlying distributions, making it suitable for a wide range of applications in various fields such as statistics, machine learning, and data science.

1.3 Thesis Overview

This thesis is organized as follows:

☐ Chapter 2: We propose a Libby-Novick Beta mixture model, which is based on a generalization of Beta distribution. As a result of having an additional shape parameter, the Libby-Novick Beta distribution provides more flexibility than conventional Beta distribution or other common distributions such as Gaussian distribution. To estimate the parameters of this novel mixture model, we applied the maximum likelihood technique. To demonstrate the robustness of our proposed model, we compared it to other alternatives. We tested this novel unsupervised model on three real and publicly available medical datasets.

☐ Chapter 3: We extend the minimum message length (MML) principle to determine the optimal number of clusters in a finite Libby-Novick Beta mixture model. Using unsupervised algorithms, such as clustering, involves determining the number of clusters that best represent the data. We have evaluated our model against three publicly available and real-world medical datasets.

☐ Chapter 4: The present study proposes a nonparametric Bayesian methodology that employs a multivariate Libby-Novick Beta mixture model to tackle clustering challenges in data analysis, such as determining the optimal number of mixture components and estimating the model's parameters. To address this problem, we extend the finite

Libby-Novick Beta mixture model (FLNBMM) into an infinite model. This extension allows us to accurately represent the data distribution by accommodating an unspecified number of mixture components. We develop a Bayesian learning strategy that employs the Markov Chain Monte Carlo technique to estimate the posterior distribution, which provides robust power and flexibility for modeling and analyzing intricate data. We evaluate the effectiveness of our proposed method on three real-world applications.

- ☐ Chapter 5: We provide a brief and clear summary of the contributions we have made.

1.4 Publications

This thesis is based on three manuscripts. Two manuscripts are accepted as conference papers and the third one is submitted.

- ☐ Chapter 2: Niloufar Samiee, Narges Manouchehri, Nizar Bouguila, "Maximum Likelihood-Based Estimation of Finite Multivariate Libby-Novick Beta Mixture Models in Medical Applications", 2023 IEEE International Conference on Industrial Technology (ICIT) [34].

- ☐ Chapter 3: Niloufar Samiee, Narges Manouchehri, Nizar Bouguila, "Finite Libby-Novick Beta Mixture Model: An MML-Based Approach", Intelligent Information and Database Systems (ACIIDS), 2023 [35].

- ☐ Chapter 4: Niloufar Samiee, Narges Manouchehri, Nizar Bouguila, "A Nonparametric Bayesian Framework for Multivariate Libby-Novick Beta Mixture Models", Control, Decision and Information Technologies (CoDIT), 2024, submitted [36].

Chapter 2

Maximum Likelihood-Based Estimation of Finite Multivariate Libby-Novick Beta Mixture Models in Medical Applications

In this chapter, we propose to develop a model called the Libby-Novick Beta Mixture Model (LNBMM), which is based on the Libby-Novick distribution as a member of the Beta distribution family. Since it has more shape parameters, it can be fitted to the data more flexibly than conventional Beta distributions. In addition to this, the estimation of model parameters is another key challenge in mixture models. To address this, we employ the maximum likelihood (ML) and Newton-Raphson methods to accurately estimate the model parameters.

2.1 Model Specification

2.1.1 Libby-Novick Beta Distribution

We will consider x_i as a random variable, such that $0 < x_i < 1$. We suppose that x_i is following a Libby-Novick Beta distribution (LNB) with parameters $a_j, b_j, \lambda_j > 0$ [37]. Based on this assumption, we can formulate the joint

density function $p(x_i \mid a_j, b_j, \lambda_j)$ as follows:

$$p(x_i \mid a_j, b_j, \lambda_j) = \frac{\lambda_j^{a_j} {x_i}^{a_j-1}(1-x_i)^{b_j-1}}{B(a_j, b_j)\{1-(1-\lambda_j)x_i\}^{a_j+b_j}} \tag{2.1}$$

where:

$$B(a_j, b_j) = \int_0^1 t^{a_j-1}(1-t)^{b_j-1}dt = \frac{\Gamma(a_j)\Gamma(b_j)}{\Gamma(a_j+b_j)} \tag{2.2}$$

As mentioned earlier, Libby-Novick Beta is an extension of Beta distribution with an extra shape parameter which can control skewness as well as kurtosis at the same time [38]. In other words, Beta distribution can be obtained by setting shape parameter λ to one. Fig. 2.1 illustrates some examples of this distribution showing how LNB can capture patterns of data with different shapes merely by varying the value of the additional shape parameter λ.

2.1.2 Finite Libby-Novick Beta Mixture Model

We assume to have a D-dimensional data point raising from a LNB distribution represented by $\vec{X}_i = \big(x_{i1}, \ldots, x_{iD}\big)$ such that $0 < x_{id} < 1$ and $d = 1, \ldots, D$. To express a finite Libby-Novick Beta mixture including M components, we have:

$$p\big(\vec{X}_i \mid \vec{\pi}, \vec{\theta}\big) = \sum_{j=1}^{M} \pi_j p\big(\vec{X}_i \mid \vec{\theta}_j\big) \tag{2.3}$$

$\vec{\theta}_j = (\vec{a}_j, \vec{b}_j, \vec{\lambda}_j)$ and π_j are the set of parameters and weight of component j^{th}, respectively, where $j = 1, \ldots, M$. $\vec{\pi} = (\pi_1, \ldots, \pi_M)$ and $\vec{\theta} = \big(\vec{\theta}_1, \ldots, \vec{\theta}_M\big)$ are the complete set of mixture parameters that we call them $\Theta = \{\vec{\pi}, \vec{\theta}\}$, such that $\sum_{j=1}^{M} \pi_j = 1$ and $\pi_j >= 0$ for $j = 1, \ldots, M$. $\vec{a} = \big(\vec{a}_1, \ldots, \vec{a}_M\big)$, $\vec{b} = \big(\vec{b}_1, \ldots, \vec{b}_M\big)$, $\vec{\lambda} = \big(\vec{\lambda}_1, \ldots, \vec{\lambda}_M\big)$ are the parameters of mixture model. $\vec{a}_j = \big(a_{j1}, \ldots, a_{jd}\big)$, $\vec{b}_j = \big(b_{j1}, \ldots, b_{jd}\big)$, $\vec{\lambda}_j = \big(\lambda_{j1}, \ldots, \lambda_{jd}\big)$ such that $a_{jd} > 0$, $b_{jd} > 0$, $\lambda_{jd} > 0$ for $d = 1, \ldots, D$. In order to model $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$ as a dataset with N D-dimensional independent and identically distributed observations, we have Libby-Novick Beta mixture model as follows:

$$p(\mathcal{X} \mid \vec{\pi}, \vec{\theta}) = \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j p(\vec{X}_i \mid \vec{\theta}_j) \right]$$
$$= \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j \prod_{d=1}^{D} \frac{\lambda_{jd}^{a_{jd}} {x_{id}}^{a_{jd}-1} (1-x_{id})^{b_{jd}-1}}{B(a_{jd}, b_{jd}) \{1-(1-\lambda_{jd})x_{id}\}^{a_{jd}+b_{jd}}} \right] \tag{2.4}$$

Four instances of this distribution are depicted in Fig. 2.2, demonstrating the adaptability of LNBMM in modelling different data patterns.

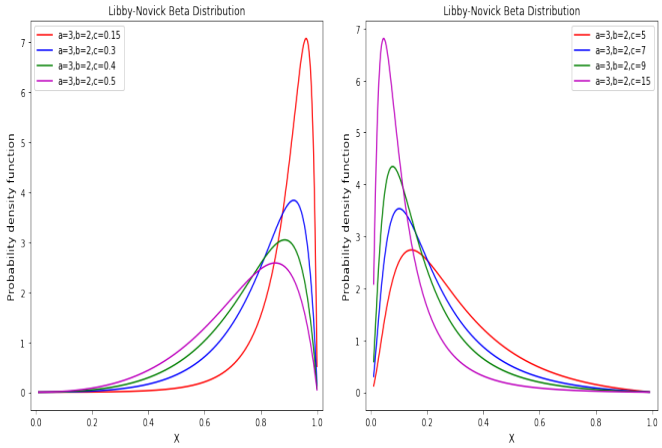

Figure 2.1: Libby-Novick Beta distribution

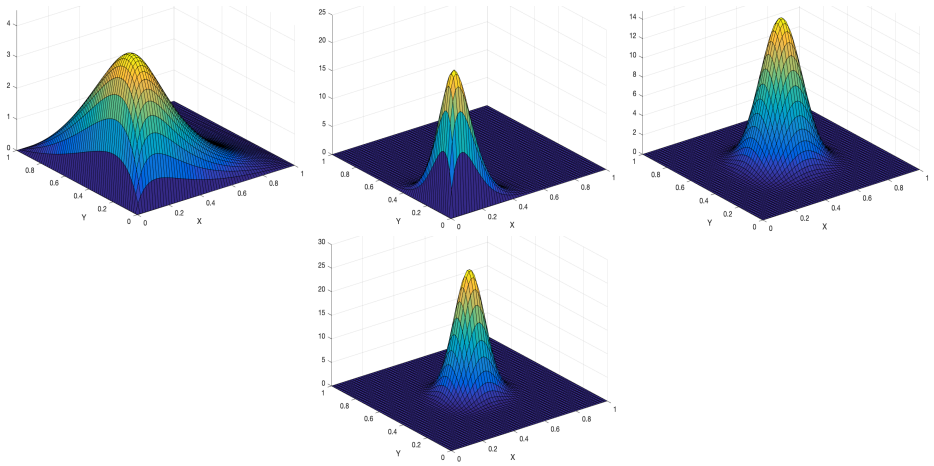

Figure 2.2: Libby-Novick Beta distribution

2.2 Model Learning

2.2.1 Maximum Likelihood and EM Algorithm

In order to estimate model parameters, we use ML methods to determine which parameters maximize the probability density function associated with the data [39]. In ML, the parameters of a mixture model are estimated to maximize the log-likelihood, which is defined as:

$$L(\Theta, \mathcal{X}) = \log p(\mathcal{X} \mid \Theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{M} \pi_j p(\vec{X}_i \mid \vec{\theta}_j) \tag{2.5}$$
$$= \sum_{i=1}^{N} \log \sum_{j=1}^{M} \pi_j \prod_{d=1}^{D} \frac{\lambda_{jd}^{a_{jd}} {x_{id}}^{a_{jd}-1}(1-x_{id})^{b_{jd}-1}}{B(a_{jd}, b_{jd})\{1-(1-\lambda_{jd})x_{id}\}^{a_{jd}+b_{jd}}}$$

As each $\vec{X}_i$ is assigned to a component j, we define a vector $\vec{Z}_i = (Z_{i1}, ..., Z_{ij})$ such that $Z_{ij} = 1$ if $\vec{X}_i$ belongs to component j, else 0 and $\sum_{j=1}^{M} Z_{ij} = 1$. For $\mathcal{X}$, we define a set of membership vectors $\mathcal{Z} = \{Z_1, \ldots, Z_N\}$.

$$Z_{ij} = \begin{cases} 1 & if \vec{X}_i \text{ in cluster j} \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

It is also important to note that each vector $\vec{X}_i$ is assigned to one of the M clusters based on its posterior probability as follows:

$$\hat{Z}_{ij} = p(j \mid \vec{X}_i, \vec{\theta}_j) = \frac{\pi_j p(\vec{X}_i, \vec{\theta}_j)}{\sum_{j=1}^{M} \pi_j p(\vec{X}_i, \vec{\theta}_j)} \tag{2.7}$$

As a result, the log-likelihood can be formulated as follows:

$$
\begin{aligned}
L(\Theta, \mathcal{Z}, \mathcal{X}) &= \sum_{j=1}^{M}\sum_{i=1}^{N} \hat{Z}_{ij}\Bigg[\log \pi_j \\
&+ \log \prod_{d=1}^{D} \frac{\lambda_{jd}^{a_{jd}} {x_{id}}^{a_{jd}-1}(1-x_{id})^{b_{jd}-1}}{B(a_{jd}, b_{jd})\{1-(1-\lambda_{jd})x_{id}\}^{a_{jd}+b_{jd}}}\Bigg] \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N} \hat{Z}_{ij}(\log \pi_j + \sum_{d=1}^{D}\Bigg[a_{jd}\log\lambda_{jd} + a_{jd}\log x_{id} \\
&- \log x_{id} + b_{jd}\log(1-x_{id}) - \log(1-x_{id}) \\
&+ \log\Gamma(a_{jd}+b_{jd}) - \log\Gamma(a_{jd}) - \log\Gamma(b_{jd}) \\
&- a_{jd}\log(1-(1-\lambda_{jd})x_{id}) - b_{jd}\log(1-(1-\lambda_{jd})x_{id})\Bigg]
\end{aligned}
\tag{2.8}
$$

We calculate the gradient of the log-likelihood with respect to the parameters in order to maximize the complete log-likelihood.

$$
\frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \Theta} = 0 \tag{2.9}
$$

2.2.2 Newton-Raphson Method

As there is no closed-form solution to (4.8), Newton-Raphson is used to update the parameters iteratively [40]. G is the gradient, which is the first derivative of $L(\Theta, \mathcal{Z}, \mathcal{X})$ with respect to the parameters and H is the Hessian matrix, which is the second and mixed derivatives of $L(\Theta, \mathcal{Z}, \mathcal{X})$ with respect to the parameters. The following steps are followed in order to update parameters:

$$
\begin{aligned}
\hat{a_j}^{new} &= {a_j}^{old} - {H_j}^{-1}G_j \\
\hat{b_j}^{new} &= {b_j}^{old} - {H_j}^{-1}G_j \\
\hat{\lambda_j}^{new} &= {\lambda_j}^{old} - {H_j}^{-1}G_j
\end{aligned}
\tag{2.10}
$$

In the next step, we calculate the derivatives with respect to a_{jd}, b_{jd} and λ_{jd}. Assuming that $\psi = \frac{\Gamma'(X)}{\Gamma(X)}$, the following equations can be obtained:

$$G_{1jd} = \frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd}} \tag{2.11}$$
$$= \sum_{i=1}^{N} \hat{Z}_{ij} \Bigg[\log \lambda_{jd} + \log x_{id} - \psi(a_{jd}) + \psi(a_{jd} + b_{jd}) - \log(1 - (1 - \lambda_{jd}) x_{id}) \Bigg]$$

$$G_{2jd} = \frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd}} \tag{2.12}$$
$$= \sum_{i=1}^{N} \hat{Z}_{ij} \Bigg[\log(1 - x_{id}) - \psi(b_{jd}) + \psi(a_{jd} + b_{jd}) - \log(1 - (1 - \lambda_{jd}) x_{id}) \Bigg]$$

$$G_{3jd} = \frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \lambda_{jd}} = \sum_{i=1}^{N} \hat{Z}_{ij} \Bigg[\frac{a_{jd}}{\lambda jd} - \frac{a_{jd} x_{id}}{(1 - (1 - \lambda_{jd}) x_{id})} - \frac{b_{jd} x_{id}}{(1 - (1 - \lambda_{jd}) x_{id})} \Bigg] \tag{2.13}$$

In order to calculate the Hessian matrix, we compute the second and mixed derivatives of the log-likelihood function.

- Derivatives with respect to a_{jd}, a_{jd}:

$$H_{a_{jd}, a_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd}^2} \tag{2.14}$$
$$= \sum_{i=1}^{N} \hat{Z}_{ij} [\psi'(a_{jd} + b_{jd}) - \psi'(a_{jd})]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd_g} \partial a_{jd_h}} = 0, d_g \neq d_h \tag{2.15}$$

- Derivatives with respect to b_{jd}, a_{jd}:

$$H_{b_{jd},a_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd} \partial a_{jd}} \tag{2.16}$$

$$= \sum_{i=1}^{N} \hat{Z}_{ij}[\psi'(a_{jd} + b_{jd})]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd_g} \partial a_{jd_h}} = 0, d_g \neq d_h \tag{2.17}$$

- Derivatives with respect to a_{jd}, λ_{jd}:

$$H_{\lambda_{jd},a_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \lambda_{jd} \partial a_{jd}} \tag{2.18}$$

$$= \sum_{i=1}^{N} \hat{Z}_{ij}[\frac{1}{\lambda jd} - \frac{x_{id}}{(1 - (1 - \lambda_{jd})x_{id})})]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \lambda_{jd_g} \partial a_{jd_h}} = 0, d_g \neq d_h \tag{2.19}$$

- Derivatives with respect to a_{jd}, b_{jd}:

$$H_{a_{jd},b_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd} \partial b_{jd}} \tag{2.20}$$

$$= \sum_{i=1}^{N} \hat{Z}_{ij}[\psi'(a_{jd} + b_{jd})]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd_g} \partial b_{jd_h}} = 0, d_g \neq d_h \tag{2.21}$$

- Derivatives with respect to b_{jd}, b_{jd}:

$$H_{b_{jd},b_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd}^2} \tag{2.22}$$

$$= \sum_{i=1}^{N} \hat{Z}_{ij}[\psi'(a_{jd} + b_{jd}) - \psi'(b_{jd})]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd_g} \partial b_{jd_h}} = 0, d_g \neq d_h \tag{2.23}$$

- Derivatives with respect to λ_{jd}, b_{jd}:

$$H_{\lambda_{jd},b_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \lambda_{jd} \partial b_{jd}} \tag{2.24}$$

$$= \sum_{i=1}^{N} \hat{Z}_{ij} [-\frac{x_{id}}{(1-(1-\lambda_{jd})x_{id})}]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \lambda_{jd_g} \partial b_{jd_h}} = 0, d_g \neq d_h \tag{2.25}$$

- Derivatives with respect to a_{jd}, λ_{jd}:

$$H_{a_{jd},\lambda_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd} \partial \lambda_{jd}} \tag{2.26}$$

$$= \sum_{i=1}^{N} \hat{Z}_{ij} [\frac{1}{\lambda_{jd}} - \frac{x_{id}}{(1-(1-\lambda_{jd})x_{id})}]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial a_{jd_g} \partial \lambda_{jd_h}} = 0, d_g \neq d_h \tag{2.27}$$

- Derivatives with respect to b_{jd}, λ_{jd}:

$$H_{b_{jd},\lambda_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd} \partial \lambda_{jd}} \tag{2.28}$$

$$= \sum_{i=1}^{N} \hat{Z}_{ij} [-\frac{x_{id}}{(1-(1-\lambda_{jd})x_{id})}]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial b_{jd_g} \partial \lambda_{jd_h}} = 0, d_g \neq d_h \tag{2.29}$$

- Derivatives with respect to $\lambda_{jd}, \lambda_{jd}$:

$$H_{\lambda_{jd},\lambda_{jd}} = \frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \lambda_{jd}^2} \tag{2.30}$$

$$= \sum_{i=1}^{N} \hat{Z}_{ij} [\frac{a_{jd} x_{id}^2}{(1-(1-\lambda_{jd})x_{id})^2} - \frac{a_{jd}}{\lambda_{jd}^2}$$

$$+ \frac{b_{jd} x_{id}^2}{(1-(1-\lambda_{jd})x_{id})^2}]$$

$$\frac{\partial^2 L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \lambda_{jd_g} \partial \lambda_{jd_h}} = 0, d_g \neq d_h \tag{2.31}$$

Our Hessian matrix is a $3D$ by $3D$ matrix as shown below:

$$H_j = \begin{bmatrix} H_{(a_{jd},a_{jd})} & H_{(a_{jd},b_{jd})} & H_{(a_{jd},p_{jd})} \\ H_{(b_{jd},a_{jd})} & H_{(b_{jd},b_{jd})} & H_{(b_{jd},p_{jd})} \\ H_{(p_{jd},a_{jd})} & H_{(p_{jd},b_{jd})} & H_{(p_{jd},p_{jd})} \end{bmatrix} \tag{2.32}$$

To estimate the values of mixing proportion we will follow this equation:

$$\pi_j = \frac{\sum_{n=1}^{N} p\big(j \mid \vec{X}_i, \vec{\theta}_j\big)}{N} \tag{2.33}$$

In order to determine the initial mixing proportions, we use K-means [41]. Our model's performance will be optimal if initialization is performed appropriately to avoid convergence to a local maximum, which cannot be guaranteed using EM.

2.2.3 Parameter Estimation Algorithm

We summarize all steps of our model in the following algorithm:

Algorithm 1 Parameter Estimation Algorithm

1. Input X and the number of clusters M.
2. Use K-Means algorithm to initialize the M clusters.
3. Initialize the parameters.
   Repeat
4. EM algorithm
   (a) E step: Compute $\hat{Z_{nj}}$.
   (b) M step: Update parameters and mixing proportions.
   (c) If $\pi_j < \epsilon$ then delete component j return to E step

   until Convergence

2.3 Experimental Results

In this part of our research, we applied our suggested algorithm to three real-world medical imaging applications, including images of lung tissue samples, histological breast tissue images, and histological malaria images. In order to apply our models to these datasets, we must extract first features from images and apply some preprocessing. For feature extraction we used Scale-invariant Feature Transform (SIFT) [42] and Bag of Visual Words (BoVW) [43]. A bag of visual words depicts images as patches and their distinctive patterns are derived from each image. For the purpose of extracting these visual features, we use SIFT as a feature detector. During preprocessing, we normalized our datasets by applying the min-max method since we suppose all input values are between (0,1):

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.34}$$

The performance of our model is compared with that of Gaussian mixture models and Beta mixture models. Our performance was evaluated using four metrics: accuracy, precision, recall, and FPR (False Positive Rate) or F1-score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.35}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.36}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.37}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.38}$$

2.3.1 Breast Tissue Analysis

Breast cancer is the most prevalent cancer among women and the main cause of cancer-related mortality globally [44]. As breast cancer cells can spread to other areas of the body if they enter the bloodstream or lymphatic system, it is very critical to detect and treat the disease early in order to increase the chance of survival. As the disease is diagnosed at a very late stage in many cases, the strategies proposed to reduce mortality and manage this disease

have not been as successful as expected. A mammogram or ultrasound is typically the first imaging tool that detects breast cancer in the majority of cases. For a precse diagnosis of a radiographic abnormality, a tissue sample is necessary. An expert pathologist evaluates tissues in order to determine whether they are benign or malignant (cancerous). Breast tissue samples can be obtained through different methods of biopsy. The tissue from the biopsy is processed in the pathology lab. Microscopically examining a biopsy or surgical specimen on a glass slide is considered histopathology, the study of disease signs. One or more stains are used to dye tissue sections under a microscope so that different components of the tissue can be visualized. Evaluation of breast cancer histology involves microscopic analysis of the chemical and cellular characteristics of cells from a suspicious tumor. On the basis of histological characteristics, benign and malignant lesions can be distinguished. Fig. 4.4 shows a few samples of breast tissues. According to both groups of images, darker, more violet-colored tissue is more likely to be cancerous than those with rose colors. But as you can see, this is not always true, so learning the model becomes more challenging. Machine learning techniques could reduce the number of false diagnoses and increase the precision of breast cancer diagnosis. We tested our method on a publicly available dataset with malignant and benign labels, each containing 500 samples [45]. We find that our proposed model performs better than GMM and BMM according to Table II.

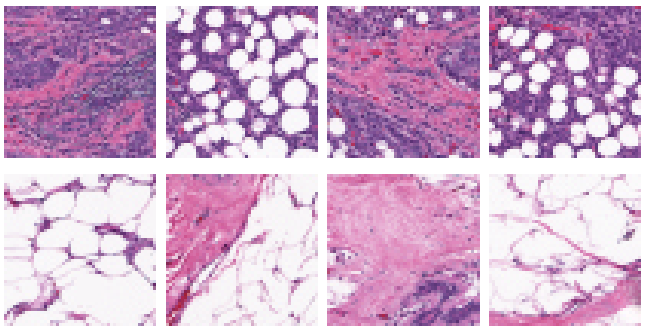

Figure 2.3: Breast tissue samples. First and second rows show benign and malignant samples.

Table 2.1: Results on Breast Tissue Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
LNBMM	88.0.72	86.54	90.01	88.24
GMM	73.7	82.11	60.6	69.73
BMM	75.04	74.05	78.2	76.01

2.3.2 Malaria Detection

Malaria is a potentially fatal disease caused by parasites transferred to humans by the bites of infected female Anopheles mosquitoes. There are measures that can be taken to prevent and treat it. The diagnosis of malaria must be made in a timely and precise manner in order to provide effective treatment [46]. In some cases, malaria may require immediate medical attention depending on its severity. The delay in diagnosing and treating malaria reduces the likelihood of survival for patients, and it causes the deaths of many people in some parts of the world. The use of microscopes is one of the most common methods of diagnosing malaria. In order to identify malaria parasites, a drop of blood can be spread on a microscope slide to form a "blood smear". Typically, a blood smear determines the number and shape of blood cells. Malaria is diagnosed by microscopists via finding malaria-infected blood cells in blood smears. There is a small clot inside infected cells, whereas uninfected cells have no clot. A few images of blood cells are shown in Fig. 4.2. A malaria parasite causes small protein nodules called knobs to appear on the surface of red blood cells, which normally have a smooth surface. It is quite challenging for humans to find positive results in a large number of smears and to verify every sample accurately. Additionally, diagnosis depends on the quality of the microscope and the experience of the technician. As a result of the reasons mentioned above, the use of a reliable machine learning technique can be very valuable in diagnosing this disease. In this research, we used a cell image dataset from NIH [46, 47], including 2000 samples, to assess our model.

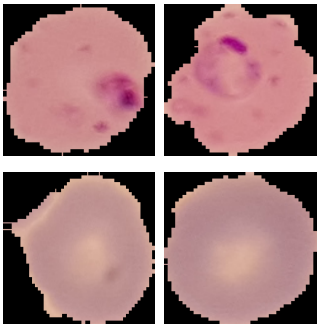

Figure 2.4: Samples of infected and uninfected cells are shown in the first and second rows, respectively.

Table 2.2: Results on Malaria Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
LNBMM	88.63	85.13	93.61	89.16
GMM	83.3	74.96	100	85.68
BMM	73.05	70.80	78.45	74.43

2.3.3 Lung Cancer Diagnosis

In the second part of our experiment, we applied our model to lung histopathological images [48]. In this dataset, there are 2500 images that include benign, adenocarcinoma, and squamous cell carcinoma tissues related to lung cancer. Fig. 4.3 shows an example of each class. As with other types of cancer, an early and accurate diagnosis is critical for the treatment of lung cancer. In Table 4.2, we demonstrate that our proposed algorithm has better performance than the GMM and BMM.

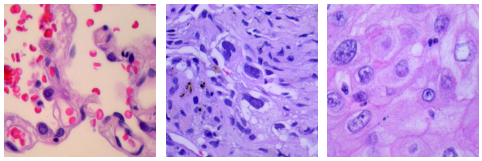

Figure 2.5: Images of three types of lung tissues, including benign ones, adenocarcinomas, and squamous cell carcinomas.

Table 2.3: Results on Lung Cancer Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
LNBMM	91.31	93.33	88.61	90.91
GMM	83.33	79.88	80.43	81.58
BMM	79.33	72.73	86.12	78.86

Chapter 3

Finite Libby-Novick Beta Mixture Model: An MML-Based Approach

In this chapter, we propose an unsupervised algorithm for learning the optimal number of clusters in a finite Libby-Novick Beta mixture model. In unsupervised learning, it is crucial to determine the number of clusters that best describes the data. By extending the minimum message length (MML) principle [49], we are able to determine the number of clusters in Libby-Novick Beta mixtures.

3.1 Model Specification

3.1.1 Finite Libby-Novick Beta Mixture Model

We assume that $\vec{X}_i = \big(x_{i1}, \ldots, x_{iD}\big)$, a D-dimensional vector such that $0 \leq x_{id} \leq 1$ and $d = 1, \ldots, D$, follows a Libby-Novick Beta mixture model [37,38]. We consider $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$ as a dataset with N independent and identically distributed D-dimensional observations. In the following equa-

tion, we have the Libby-Novick Beta mixture model:

$$p(\mathcal{X} \mid \vec{\pi}, \vec{\theta}) = \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j p(\vec{X}_i \mid \vec{\theta}_j) \right]$$
$$= \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j \prod_{d=1}^{D} \frac{\lambda_{jd}^{a_{jd}} x_{id}^{a_{jd}-1} (1-x_{id})^{b_{jd}-1}}{B(a_{jd}, b_{jd})\{1-(1-\lambda_{jd})x_{id}\}^{a_{jd}+b_{jd}}} \right] \quad (3.1)$$

The parameters of component j^{th} and its weight are given by $\vec{\theta}_j = (\vec{a}_j, \vec{b}_j, \vec{\lambda}_j)$ and π_j, where $j = 1, \ldots, M$. $\vec{\pi} = (\pi_1, \ldots, \pi_M)$, $\vec{\theta} = (\vec{\theta}_1, \ldots, \vec{\theta}_M)$, and $\Theta = \{\vec{\pi}, \vec{\theta}\}$, the complete set of mixture parameters that $\sum_{j=1}^{M} \pi_j = 1$ and $\pi_j >= 0$ for $j = 1, \ldots, M$. $\vec{a} = (\vec{a}_1, \ldots, \vec{a}_M)$, $\vec{b} = (\vec{b}_1, \ldots, \vec{b}_M)$, $\vec{\lambda} = (\vec{\lambda}_1, \ldots, \vec{\lambda}_M)$ are the parameters of mixture model, such that $a_{jd} > 0$, $b_{jd} > 0$, $\lambda_{jd} > 0$ for $d = 1, \ldots, D$. In the following subsection, we give a brief summary of the maximum likelihood approach that we developed in [34] to estimate the parameters of the proposed mixture model. The estimation of these parameters is an important step towards developing the MML criterion.

3.1.2 Maximum Likelihood and EM Algorithm

As part of the estimation process, our model parameters are calculated using ML estimation within an EM framework, which determine the parameters that maximize the model's likelihood function. In fact, ML estimates mixture model parameters in order to maximize the log-likelihood [34]. We define a vector $\vec{Z}_i = (Z_{i1}, ..., Z_{ij})$ such that $Z_{ij} = 1$ if $\vec{X}_i$ belongs to component j and 0 otherwise and $\sum_{j=1}^{M} Z_{ij} = 1$. As a result, we define a set of membership vectors $\mathcal{Z} = \{Z_1, \ldots, Z_N\}$ for the set $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$. The posterior probability of each vector $\vec{X}_i$ determines its assignment to one of the M clusters, as follows:

$$\hat{Z}_{ij} = p(j \mid \vec{X}_i, \vec{\theta}_j) = \frac{\pi_j p(\vec{X}_i, \vec{\theta}_j)}{\sum_{j=1}^{M} \pi_j p(\vec{X}_i, \vec{\theta}_j)} \quad (3.2)$$

This leads to the following formulation for log-likelihood:

$$
\begin{aligned}
L(\Theta, \mathcal{Z}, \mathcal{X}) &= \sum_{j=1}^{M}\sum_{i=1}^{N}\hat{Z}_{ij}\Bigg[\log \pi_j \\
&+ \log\prod_{d=1}^{D}\frac{\lambda_{jd}^{a_{jd}}{x_{id}}^{a_{jd}-1}(1-x_{id})^{b_{jd}-1}}{B(a_{jd},b_{jd})\{1-(1-\lambda_{jd})x_{id}\}^{a_{jd}+b_{jd}}}\Bigg] \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}\hat{Z}_{ij}(\log\pi_j + \sum_{d=1}^{D}\Bigg[a_{jd}\log\lambda_{jd} + a_{jd}\log x_{id} \\
&- \log x_{id} + b_{jd}\log(1-x_{id}) - \log(1-x_{id}) \\
&+ \log\Gamma(a_{jd}+b_{jd}) - \log\Gamma(a_{jd}) - \log\Gamma(b_{jd}) \\
&- a_{jd}\log(1-(1-\lambda_{jd})x_{id}) - b_{jd}\log(1-(1-\lambda_{jd})x_{id})\Bigg]
\end{aligned}
\tag{3.3}
$$

In order to maximize the complete log-likelihood, we compute the gradient of log-likelihood with respect to the parameters. In this case, Newton-Raphson is used as an iterative method to update the parameters [34], as there is no closed-form solution to following equation:

$$
\frac{\partial L(\Theta, \mathcal{Z}, \mathcal{X})}{\partial \Theta} = 0 \tag{3.4}
$$

For more details about the parameters estimation, the reader is referred to [34].

3.1.3 The MML Criterion For a Finite Libby-Novick Beta Mixture

As a model selection technique, MML (minimum message length) is used in this section. According to information theory, the optimal number of clusters transmits data from sender to receiver efficiently with the least amount of information. As a result, MML can be defined as follows for a mixture of distributions:

$$
\text{MML} = -\log(\frac{h(\Theta)p(\mathcal{X} \mid \Theta)}{\sqrt{\mid F(\Theta) \mid}}) + N_p(-\frac{1}{2}\log(12) + \frac{1}{2}) \tag{3.5}
$$

where $h(\Theta)$ is prior probability distribution, $p(\mathcal{X} \mid \Theta)$ is the complete data log-likelihood, $F(\Theta)$ is the expected Fisher information matrix computed by taking the second derivative of the negative log-likelihood, and $|F(\Theta)|$ is its determinant. N_p represents the number of free parameters and equals $(M(2D+1))-1$.

3.1.4 Fisher Information for Libby-Novick Beta mixture model

As the expected value of the negative of the Hessian matrix, the Fisher matrix, also known as the curvature matrix, describes the curve of the likelihood function around its maximum. Since MML is based on a Hessian matrices, it takes on a sophisticated analytical form that is difficult to reproduce. As a result, we will approximate this matrix using these two following assumptions: First, it is important to keep in mind that $\vec{\theta}$ and the vector $\vec{\pi}$ are independent, as one's preconceived notions about the value of the mixing parameter vector $\vec{\pi}$ do not typically influence one's notions about $\vec{\theta}$. Moreover, we presume that the $\vec{\theta}$ components are also independent. Fisher information matrix can be calculated after clustering data vectors according to a mixture model. A Fisher information matrix has the following determinant:

$$|F(\Theta)| = |F(\vec{\pi})| \prod_{j=1}^{M} |F(\vec{\theta_j})| \tag{3.6}$$

$|F(\vec{\pi})|$ is the determinant of Fisher information of mixing parameters π_j and $|F(\vec{\theta_j})|$ is the determinant of Fisher information with regard to the vector $\vec{\theta_j} = (\vec{a}_j, \vec{b}_j, \vec{p}_j)$ of a single Libby-Novick Beta distribution. As a result, we can compute the Fisher information matrix determinant by assuming a generalized Bernoulli process where there are M possible results for M clusters for each trial as follows:

$$|F(\vec{\pi})| = \frac{N^{M-1}}{\prod_{j=1}^{M} \pi_j} \tag{3.7}$$

The Fisher information for our mixture is as follows:

$$\log(|F(\Theta)|) = (M-1)\log(N) - \sum_{j=1}^{M} \log(\pi_j) + \sum_{j=1}^{M} \log(|F(\vec{\theta_j})|) \tag{3.8}$$

3.1.5 Determinant of the Fisher information

Assuming the mixture model procedure is followed, $X_j = (\vec{X}_t, \ldots, \vec{X}_{t+n_j-1})$ samples of data are allocated to cluster jth, such that $t \leq N$ and n_j is the number of samples assigned to cluster j.

$$\begin{aligned} -\log p(\mathcal{X} \mid \Theta) &= -\log(\prod_{n=t}^{t+n_j-1} p(\vec{X} \mid \vec{\theta}_M)) \\ &= -(\sum_{n=t}^{t+n_j-1} \log p(\vec{X} \mid \vec{\theta}_M)) \end{aligned} \tag{3.9}$$

$F(\vec{\theta_j})$ is defined as the negative of the second derivative of complete log-likelihood. In accordance with the parameters $a_{jd}, b_{jd}, \lambda_{jd}$, we calculate the second and mixed derivatives:

- Derivatives with respect to a_{jd}, a_{jd}:

$$F_{a_{jd},a_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd}^2} = -n_j(\psi'(a_{jd} + b_{jd}) - \psi'(a_{jd})) \tag{3.10}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd_s} \partial a_{jd_t}} = 0, d_s \neq d_t \tag{3.11}$$

- Derivatives with respect to a_{jd}, b_{jd}:

$$F_{a_{jd},b_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd} \partial b_{jd}} = -n_j(\psi'(a_{jd} + b_{jd})) \tag{3.12}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd_s} \partial b_{jd_t}} = 0, d_s \neq d_t \tag{3.13}$$

- Derivatives with respect to a_{jd}, λ_{jd}:

$$F_{a_{jd},\lambda_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd} \partial \lambda_{jd}} = -\sum_{i=1}^{N} \hat{Z}_{ij}[\frac{1}{\lambda jd} - \frac{x_{id}}{(1-(1-\lambda_{jd})x_{id})})] \tag{3.14}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial a_{jd_s} \partial \lambda_{jd_t}} = 0, d_s \neq d_t \tag{3.15}$$

- Derivatives with respect to b_{jd}, a_{jd}:

$$F_{b_{jd},a_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd} \partial a_{jd}} = -n_j(\psi'(a_{jd} + b_{jd})) \tag{3.16}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd_s} \partial a_{jd_t}} = 0, d_s \neq d_t \tag{3.17}$$

- Derivatives with respect to b_{jd}, b_{jd}:

$$F_{b_{jd},b_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd}^2} = -n_j(\psi'(a_{jd} + b_{jd}) - \psi'(b_{jd})) \tag{3.18}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd_s} \partial b_{jd_t}} = 0, d_s \neq d_t \tag{3.19}$$

- Derivatives with respect to b_{jd}, λ_{jd}:

$$F_{b_{jd},\lambda_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd} \partial \lambda_{jd}} = -\sum_{i=1}^{N} \hat{Z}_{ij}[-\frac{x_{id}}{(1-(1-\lambda_{jd})x_{id})}] \tag{3.20}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial b_{jd_s} \partial \lambda_{jd_t}} = 0, d_s \neq d_t \tag{3.21}$$

- Derivatives with respect to λ_{jd}, a_{jd}:

$$F_{\lambda_{jd},a_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial \lambda_{jd} \partial a_{jd}} = -\sum_{i=1}^{N} \hat{Z}_{ij}[\frac{1}{\lambda_{jd}} - \frac{x_{id}}{(1-(1-\lambda_{jd})x_{id})}] \tag{3.22}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial \lambda_{jd_s} \partial a_{jd_t}} = 0, d_s \neq d_t \tag{3.23}$$

- Derivatives with respect to λ_{jd}, b_{jd}:

$$F_{\lambda_{jd},b_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial \lambda_{jd} \partial b_{jd}} = -\sum_{i=1}^{N} \hat{Z}_{ij}[-\frac{x_{id}}{(1-(1-\lambda_{jd})x_{id})}] \tag{3.24}$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial \lambda_{jd_s} \partial b_{jd_t}} = 0, d_s \neq d_t \tag{3.25}$$

- Derivatives with respect to $\lambda_{jd}, \lambda_{jd}$:

$$F_{\lambda_{jd},\lambda_{jd}} = -\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial \lambda_{jd}^2} \tag{3.26}$$

$$= -\sum_{i=1}^{N} \hat{Z}_{ij}[\frac{a_{jd}x_{id}^2}{(1-(1-\lambda_{jd})x_{id})^2} - \frac{a_{jd}}{\lambda_{jd}^2} + \frac{b_{jd}x_{id}^2}{(1-(1-\lambda_{jd})x_{id})^2}]$$

$$-\frac{\partial^2 \log p(\mathcal{X} \mid \Theta)}{\partial \lambda_{jd_s} \partial \lambda_{jd_t}} = 0, d_s \neq d_t \tag{3.27}$$

The $F(\vec{\theta_j})$ is a $3D$ by $3D$ matrix as follows:

$$F_j = \begin{bmatrix} F_{(a_{jd},a_{jd})} & F_{(a_{jd},b_{jd})} & F_{(a_{jd},\lambda_{jd})} \\ F_{(b_{jd},a_{jd})} & F_{(b_{jd},b_{jd})} & F_{(b_{jd},\lambda_{jd})} \\ F_{(\lambda_{jd},a_{jd})} & F_{(\lambda_{jd},b_{jd})} & F_{(\lambda_{jd},\lambda_{jd})} \end{bmatrix} \tag{3.28}$$

3.1.6 Prior distribution

We must choose the model's parameters' prior distribution $h(\Theta)$ in order to calculate the MML criterion. We define $h(\Theta)$ as follows:

$$h(\Theta) = h(\vec{\pi})h(\vec{a})h(\vec{b})h(\vec{\lambda}) \tag{3.29}$$

For modelling proportional vectors, we assume a Dirichlet distribution for $h(\vec{\pi})$ where $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_M)$:

$$h(\pi_1, \pi_2, \ldots, \pi_M) = \frac{\Gamma(\sum_{j=1}^{M} \eta_j)}{\prod_{j=1}^{M} \Gamma(\eta_j)} \prod_{j=1}^{M} \pi_j^{\eta_j - 1} \tag{3.30}$$

By calculating a uniform prior for the parameter $\eta, (\eta_1 = 1, ..., \eta_M = 1)$, we can simplify (3.30) as follows:

$$h(\vec{\pi}) = (M-1)! \tag{3.31}$$

We assume that dimensions are independent, so we have:

$$h(\vec{a}) = \prod_{j=1}^{M} \prod_{d=1}^{D} h(a_{jd}) \tag{3.32}$$

The assumption is that we do not have any prior information regarding parameter a_{jd} in this case. Consequently, we use a simple uniform prior in accordance with Ockham's razor, which has shown to be effective in producing effective results, to ensure that its effect on the posterior is minimal [50]. $h(b_{jd})$ and $h(\lambda_{jd})$ will be chosen in the same way:

$$h(a_{jd}) = e^{-6}\frac{a_{jd}}{||a_j||}, h(b_{jd}) = e^{-6}\frac{b_{jd}}{||b_j||}, h(\lambda_{jd}) = e^{-6}\frac{\lambda_{jd}}{||\lambda_j||} \tag{3.33}$$

Log of prior is provided by:

$$\begin{aligned} \log(h(\Theta)) = &-D\sum_{j=1}^{M}\log(||a_j||) + \sum_{j=1}^{M}\sum_{d=1}^{D}\log(a_{jd}) \\ &- D\sum_{j=1}^{M}\log(||b_j||) + \sum_{j=1}^{M}\sum_{d=1}^{D}\log(b_{jd}) \\ &- D\sum_{j=1}^{M}\log(||\lambda_j||) + \sum_{j=1}^{M}\sum_{d=1}^{D}\log(\lambda_{jd}) + \sum_{j=1}^{M-1}\log(j) - 18MD \end{aligned} \tag{3.34}$$

3.1.7 Full Learning Algorithm

Here is a summary of all the steps in our method:

Algorithm 2 Full Learning Algorithm

1. Input $\mathcal{X}$ and the number of clusters M.
2. Use K-Means algorithm to initialize the M clusters.
3. Initialize the parameters.
   Repeat
4. EM algorithm [34]
5. MML
   (a) Calculate the criterion of MML(M).
   (b) Find the optimal M^* i.e. $M^* = argmin_M$ MML(M).

3.2 Experimental Results

In this part of our research, we begin by extracting features from images and preprocessing our data. We employed Scale-invariant Feature Transform (SIFT) and Bag of Visual Words (BoVW) as feature extraction methods [42, 43]. Using a bag of visual words, images are presented as patches, with distinctive patterns derived from each one. SIFT is used as a feature detector for the extraction of these visual characteristics. Based on the assumption that all input values fall between 0 and 1, we normalized our datasets using the min-max method:

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.35}$$

In the following, three real-world medical imaging applications were used to evaluate our algorithm, including images of lung tissue samples, histological malaria images, and histological images of breast tissue.

3.2.1 Malaria Detection

Malaria is a parasitic disease. Individuals can become infected with a parasite when bitten by a mosquito. Using a microscope is the most common method

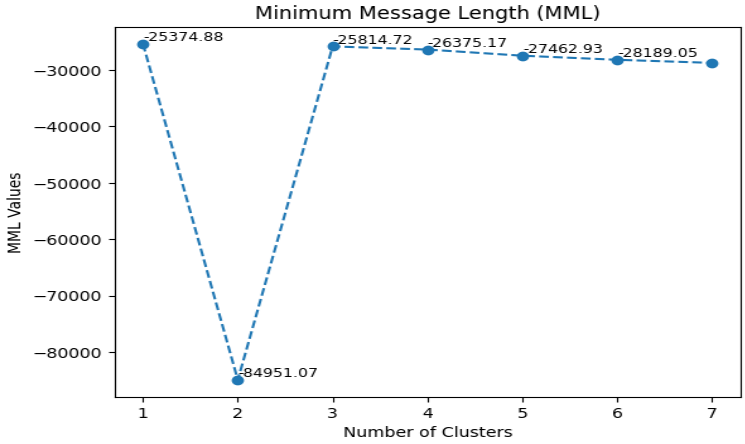

Figure 3.1: Plot of message length for the Malaria detection dataset. Clusters are represented on the X-axis, while message length is represented on the Y-axis.

of diagnosing malaria. Microscopists diagnose malaria by identifying malaria-infected blood cells in blood smears. Unlike uninfected cells, infected cells have a small clot. Humans find it difficult to detect positive results among a large number of smears and to verify each sample precisely. For this reason, using a dependable machine learning technique to diagnose this disease can be extremely helpful. As part of this study, we evaluated our model by analyzing images of 2,000 cells from the National Institutes of Health and determined the optimal number of clusters for modeling this dataset [46, 47]. Our algorithm was able to determine the optimal number of clusters as shown in Fig.3.1.

3.2.2 Breast Tissue Analysis

In the world, breast cancer is the leading cause of cancer-related death. According to current estimates, 12.9% of American women will suffer from breast cancer during their lifetimes [51]. In order to increase the chance of survival, it is essential to identify breast cancer cells at an early stage and start treatment. During the diagnostic process, an expert pathologist evaluates tissues and identifies whether they are benign or malignant (cancerous) based on their histological characteristics. It is important to note that cancer diagnosis by specialists is not error-free. Therefore, machine learning tech-

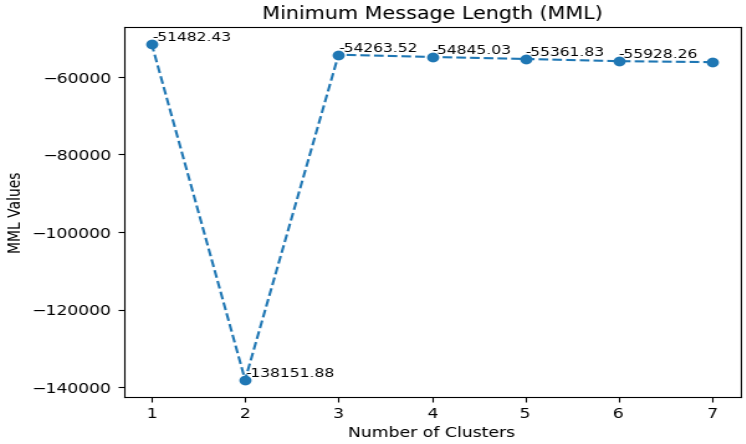

Figure 3.2: Plot of message length for the breast tissue dataset. Clusters are represented on the X-axis, while message length is represented on the Y-axis.

niques could be used to reduce the number of erroneous diagnoses as well as improve accuracy in breast cancer diagnosis. A publicly available dataset containing 500 samples, each with malignant and benign labels was used to evaluate our algorithm [45]. Fig.4.4 illustrates how our algorithm was able to determine the optimal number of clusters.

3.2.3 Lung Cancer Diagnosis

Considering that smoking is one of the major causes of lung cancer, we are seeing a large number of lung cancer cases around the globe with the increase in the number of smokers. Smoking is estimated to be responsible for approximately 80% of all lung cancer deaths [52]. This part of the paper describes how our model was applied to lung histopathological images. There are 2500 images in this dataset of lung cancer tissues including benign, adenocarcinoma, and squamous cell carcinoma [48].An example of each cluster is shown in Fig. 4.3. A timely and accurate diagnosis of lung cancer is crucial to its successful treatment, as with any other type of cancer. We were able to determine the optimal number of clusters using our algorithm, which can be seen in Fig.4.3.

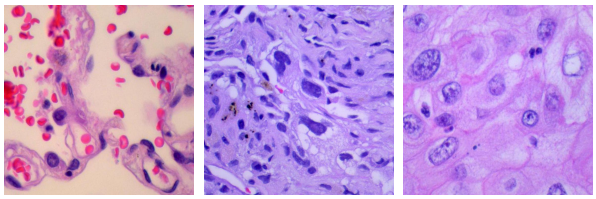

Figure 3.3: An illustration of benign lung tissue, adenocarcinoma, and squamous cell carcinoma.

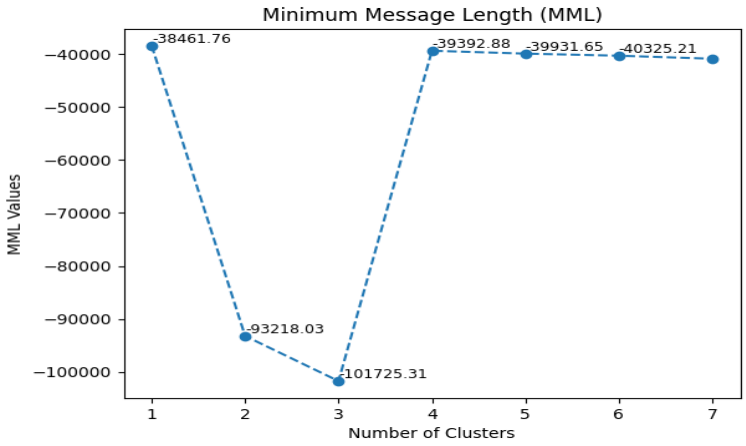

Figure 3.4: Plot of message length for lung cancer dataset. Clusters are represented on the X-axis, while message length is represented on the Y-axis.

Chapter 4

A Nonparametric Bayesian Framework for Multivariate Libby-Novick Beta Mixture Models

This chapter presents a nonparametric Bayesian approach based on a multivariate Libby-Novick Beta mixture model to address clustering challenges. Using mixture models can be challenging due to the difficulty of determining the optimal number of mixture components. To address this issue, we expand the finite Libby-Novick Beta mixture model (FLNBMM) to an infinite model. This enables us to accurately represent the data distribution by accommodating an unspecified number of mixture components. We develop a Bayesian learning strategy that uses the Markov Chain Monte Carlo technique to estimate the posterior distribution, which provides strong power and flexibility for modeling and analyzing complicated data.

4.1 Model Specification

4.1.1 Libby-Novick Beta Distribution

Let us consider a random variable x_i, $0 < x_i < 1$, that follows a Libby-Novick Beta distribution (LNB) with parameters $a_j, b_j, \lambda_j > 0$ [37]. The joint density

function of this observation $p(x_i \mid a_j, b_j, \lambda_j)$ is formulated as the following:

$$p(x_i \mid a_j, b_j, \lambda_j) = \frac{\lambda_j^{a_j} {x_i}^{a_j-1}(1-x_i)^{b_j-1}}{B(a_j, b_j)\{1-(1-\lambda_j)x_i\}^{a_j+b_j}} \tag{4.1}$$

where:

$$B(a_j, b_j) = \int_0^1 t^{a_j-1}(1-t)^{b_j-1}dt = \frac{\Gamma(a_j)\Gamma(b_j)}{\Gamma(a_j+b_j)} \tag{4.2}$$

As previously mentioned, LNB is a generalization of the Beta distribution with one additional shape parameter that controls both the skewness and kurtosis simultaneously [38]. By setting the shape parameter λ to 1, we can get the Beta distribution.

4.1.2 Finite Libby-Novick Beta Mixture Model

We can formulate a finite Libby-Novick Beta mixture with M components by assuming that we have a D-dimensional data point, following a sum of weighted LNB distributions, indicated by $\vec{X}_i = \big(x_{i1}, \ldots, x_{iD}\big)$ where $0 \leq x_{id} \leq 1$ and $d = 1, \ldots, D$. This gives us:

$$p\big(\vec{X}_i \mid \vec{\pi}, \vec{\theta}\big) = \sum_{j=1}^{M} \pi_j p\big(\vec{X}_i \mid \vec{\theta}_j\big) \tag{4.3}$$

π_j and $\vec{\theta}_j = (\vec{a}_j, \vec{b}_j, \vec{\lambda}_j)$ are the weight and set of parameters of component j^{th}, respectively where $j = 1, \ldots, M$. $\vec{\pi} = (\pi_1, \ldots, \pi_M)$ and $\vec{\theta} = \big(\vec{\theta}_1, \ldots, \vec{\theta}_M\big)$ are the complete set of mixture parameters such that $\sum_{j=1}^{M} \pi_j = 1$ and $\pi_j >= 0$ for $j = 1, \ldots, M$. $\vec{a} = \big(\vec{a}_1, \ldots, \vec{a}_M\big)$, $\vec{b} = \big(\vec{b}_1, \ldots, \vec{b}_M\big)$, $\vec{\lambda} = \big(\vec{\lambda}_1, \ldots, \vec{\lambda}_M\big)$ are the parameters of mixture model. $\vec{a}_j = \big(a_{j1}, \ldots, a_{jD}\big)$, $\vec{b}_j = \big(b_{j1}, \ldots, b_{jD}\big)$, $\vec{\lambda}_j = \big(\lambda_{j1}, \ldots, \lambda_{jD}\big)$ such that $a_{jd} > 0$, $b_{jd} > 0$, $\lambda_{jd} > 0$ for $d = 1, \ldots, D$. A few examples of this distribution can be seen in Fig. 4.1, which illustrates the capability of LNB to capture data patterns of various shapes.

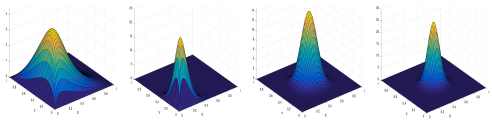

Figure 4.1: Libby-Novick Beta distribution

We formulate LNB mixture model (LNBMM) as below in order to model $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$ as a dataset with N D-dimensional independent and identically distributed observations.

$$p(\mathcal{X} \mid \vec{\pi}, \vec{\theta}) = \prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j p(\vec{X}_i \mid \vec{\theta}_j) \right] = \tag{4.4}$$

$$\prod_{i=1}^{N} \left[\sum_{j=1}^{M} \pi_j \prod_{d=1}^{D} \frac{\lambda_{jd}^{a_{jd}} {x_{id}}^{a_{jd}-1} (1 - x_{id})^{b_{jd}-1}}{B(a_{jd}, b_{jd}) \{1 - (1 - \lambda_{jd}) x_{id}\}^{a_{jd}+b_{jd}}} \right]$$

4.2 Model Learning

4.2.1 Bayesian Learning Framework

The process of implementing mixture models involves dealing with two main challenges. Firstly, there is the issue of accurately estimating the model parameters, which can be a complex task. Secondly, it is important to determine the optimal number of components for the model, as this can significantly impact the accuracy and effectiveness of the model [53]. Choosing the appropriate number of components requires careful consideration and analysis to ensure that the resulting model is both reliable and useful. A popular method for estimating model parameters in statistical modeling is to employ the EM algorithm. This algorithm maximizes the likelihood function of the data by introducing a latent indicator variable $\mathcal{Z} = \{\vec{Z}_1, \ldots, \vec{Z}_N\}$. For each observation $\vec{X}_i$, $\vec{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ indicates which component it belongs to [54], [55]. Therefore, if $\vec{X}_i$ is the most likely to belong to cluster j, $Z_{ij} = 1$ and for other clusters $Z_{ij} = 0$. In view of the membership vectors for $\mathcal{X}$, we then have the complete form of data as $(\mathcal{X}, \mathcal{Z})$, which follows $p(\mathcal{X}, \mathcal{Z} \mid \Theta)$. The symbol $\Theta = (\vec{\theta}, \vec{\pi})$ indicates the complete set of parameters. As a result, the likelihood function of complete data can be defined as follows:

$$p(\mathcal{X}, \mathcal{Z} \mid \Theta) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left(p(\vec{X}_i \mid \vec{\theta}_j) \pi_j \right)^{Z_{ij}} \tag{4.5}$$

In the context of learning model parameters, we propose employing a Bayesian framework for LBNMM, given its exceptional properties and benefits over likelihood-based approaches. Bayesian inference aims to estimate the distribution over the model parameters, as opposed to a single set of parameters,

which distinguishes it from likelihood-based methods [21,31,56–58]. This fundamental difference in approach can have significant implications for model performance and reliability [59]. In the Bayesian learning framework, estimating the posterior distribution of the mixture model is done by Markov Chain Monte Carlo (MCMC) and by employing Gibbs sampling which is the most commonly used simulation technique. Using this method, each parameter is updated based on its conditional posterior distribution. In the Bayesian framework, determining posteriors is the most important step. It is necessary to define a prior distribution for each parameter to determine our conditional distribution's posterior distribution. Given that the parameters of the model are all positive, we suppose that all of the priors are generated using Gamma distributions. For the Gamma distribution, we use $\vec{u}, \vec{v}, \vec{r}, \vec{s}, \vec{f}, \vec{g}$ which are all positive hyper-parameters. In this regard, the prior distribution can be formulated as follows:

$$p(\vec{\theta} \mid \vec{u}, \vec{v}, \vec{r}, \vec{s}, \vec{f}, \vec{g}) = \prod_{j=1}^{M} \prod_{d=1}^{D} p\left(a_{jd} \mid u_{jd}, v_{jd}\right) \\ p\left(b_{jd} \mid s_{jd}, r_{jd}\right)\ p\left(\lambda_{jd} \mid f_{jd}, g_{jd}\right) \tag{4.6}$$

To calculate further Bayesian inference, we conditioned our likelihood on $\mathcal{Z}$ as follows:

$$p(\mathcal{X} \mid \mathcal{Z}, \vec{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left(\pi_j \prod_{d=1}^{D} p(x_{id} \mid \theta_{jd}) \right)^{Z_{ij}} \tag{4.7}$$

Following the determination of the priors and likelihoods, the conditional posterior distributions can be calculated as follows:

$$p(a_{jd} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}) \propto p\left(a_{jd} \mid u_{jd}, v_{jd}\right) p(\mathcal{X} \mid \mathcal{Z}, \vec{\theta}) \\ p(b_{jd} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}) \propto p\left(b_{jd} \mid s_{jd}, r_{jd}\right) p(\mathcal{X} \mid \mathcal{Z}, \vec{\theta}) \\ p(\lambda_{jd} \mid \mathcal{Z}, \mathcal{X}, \vec{\theta}) \propto p\left(\lambda_{jd} \mid f_{jd}, g_{jd}\right) p(\mathcal{X} \mid \mathcal{Z}, \vec{\theta}) \tag{4.8}$$

4.2.2 Infinite Multivariate Libby-Novick Beta Mixture Model

Designing mixture models involves selecting an appropriate number of clusters M to describe the data accurately. However, choosing an appropriate

number of clusters M can be challenging. Consequently, approaches based on finite mixtures have a significant limitation in determining the number of components in advance. In order to overcome this limitation, nonparametric Bayesian approaches have been proposed, which can automatically estimate the number of clusters and increase it to infinity based on the specific choice of priors for mixing weights.
As previously explained in the finite mixture model, each vector $\vec{X}_i$ can be derived from one of M LNB distributions. In the following, we assume that the Dirichlet process of LNB distributions will be able to model our observations. A Dirichlet process mixture model is then illustrated, along with its ability to create or remove clusters. For $\vec{\pi}$, the mixing weights coefficients, we know that it is defined on $(\pi_1, \ldots, \pi_M) : \sum_{j=1}^{M-1} \pi_j < 1$, then an option to consider as a prior is a symmetric Dirichlet distribution with a concentration parameter $\frac{\eta}{M}$.

$$p(\vec{\pi} \mid \eta) = \frac{\Gamma(\eta)}{\prod_{j=1}^{M} \Gamma(\frac{\eta}{M})} \prod_{j=1}^{M} \pi_j^{\frac{\eta}{M}-1} \tag{4.9}$$

we also have:

$$p(\mathcal{Z} \mid \vec{\pi}) = \prod_{j=1}^{M} {\pi_j}^{n_j} \tag{4.10}$$

where $n_j = \sum_{i=1}^{N} \mathbb{I}_{Z_{ij}=1}$ represents the number of elements in cluster j. Due to the fact that Dirichlet is a conjugate prior to the multinomial, it will be possible to integrate the mixing proportions $\vec{\pi}$ to obtain the prior for $\mathcal{Z}$:

$$\begin{aligned} p(\mathcal{Z} \mid \eta) &= \int_{\vec{\pi}} p(Z \mid \vec{\pi}) p(\vec{\pi} \mid \eta) d\vec{\pi} \\ &= \frac{\Gamma(\eta)}{\Gamma(N+\eta)} \prod_{j=1}^{M} \frac{\Gamma(\frac{\eta}{M} + n_j)}{\Gamma(\frac{\eta}{M})} \end{aligned} \tag{4.11}$$

By combining all (4.9) to (4.11), (4.12) can be written as follows:

$$p(\vec{\pi} \mid \mathcal{Z}, \eta) = \frac{\Gamma(\eta + N)}{\prod_{j=1}^{M} \Gamma(\frac{\eta}{M} + n_j)} \prod_{j=1}^{M} \pi_j^{n_j + \frac{\eta}{M} - 1} \tag{4.12}$$

which illustrates a Dirichlet distribution with parameter $(n_1 + \frac{\eta}{M}, \ldots, n_M + \frac{\eta}{M})$. The conditional prior for one indicator is outlined in [60]:

$$p(Z_{ij} = 1 \mid \eta, \mathcal{Z}_{-i}) = \frac{n_{-ij} + \frac{\eta}{M}}{N - 1 + \eta} \tag{4.13}$$

where $\mathcal{Z}_{-i} = \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_N\}$, $n_{-i,j}$ is the number of observations excluding $\vec{X}_i$ in component j. After this, the conditional posterior is calculated by multiplying the prior (4.13) by the likelihood of $\vec{X}_i$.
As we have mentioned, an essential task in adopting mixture models is to select the complexity of the model. In this section we address this problem by considering $M \to \infty$ in (4.13) which provides the following limits [61]:

$$p(Z_{ij} = 1 \mid \eta; \mathcal{Z}_{-i}) = \begin{cases} \frac{n_{-i,j}}{N-1+\eta} & \text{if } n_{-i,j} > 0 \quad (j \in \mathcal{R}) \\ \frac{\eta}{N-1+\eta} & \text{if } n_{-i,j} = 0 \quad (j \in \mathcal{U}) \end{cases} \tag{4.14}$$

where $\mathcal{R}$ and $\mathcal{U}$ denote the sets of represented and unrepresented clusters, respectively. According to this equation, each observation has a certain probability of being assigned to either a represented component or one that is unrepresented. In the case of a represented component, the conditional prior is determined based on the number of observations already assigned to the cluster. However, for a new component (unrepresented), the conditional prior is only proportional to η and N. We can determine the conditional posteriors based on the conditional priors in (4.14) as follows [60], [62]:

$$p(Z_{ij} = 1 \mid \vec{\theta}_j, \eta; \mathcal{Z}_{-i}) = \begin{cases} \frac{n_{-i,j}}{N-1+\eta} p(\vec{X}_i \mid \vec{\theta}_j, Z_i) & \text{if } j \in \mathcal{R} \\ \int \frac{\eta p(\vec{X}_i \mid \vec{\theta}_j, Z_i) p(\vec{\theta}_j)}{N-1+\eta} d\vec{\theta}_j & \text{if } j \in \mathcal{U} \end{cases} \tag{4.15}$$

The equation that describes a Dirichlet process mixture of LNB distributions indicates the presence of an empty cluster whenever an observation is assigned to generate accordingly. This reinforces the idea of an infinite mixture model. If all observations are assigned to other clusters during sampling iterations, the represented cluster becomes empty and transforms into an unrepresented cluster. The model proposed here is comparable to a Chinese restaurant process, featuring a restaurant with an infinite number of tables. In this process, a restaurant boasting infinitely many tables is seen as a mixture of components. At the outset of the process, the first customer (i.e. data observation) takes up the initial table. As the process nears its end, the

second-to-last customer may either select the first unoccupied table or opt for an occupied one with a probability dependent on the number of people already seated there [60].

4.2.3 Complete Algorithm

The posterior distribution for infinite LNBMM is estimated using MCMC methods. These statistics-based sampling methods can be used to generate samples from complex distributions appropriately. In order to cope with the intractable forms of posterior distributions in mixture models, Gibbs sampling and Metropolis-Hastings (M-H) MCMC techniques have been employed in our study [59, 63]. First, during the initialization step, we assume that all observations belong to the same cluster. In the following step, we update the number of represented components based on the previous step, which is the generation of the $\vec{Z}_i$. Accordingly, when a sample is assigned to an unrepresented cluster, M increases by one, and if a component becomes empty during the iterations, M decreases by one [64]. It should be noted that, to obtain a sample of the vector $\vec{Z}_i$, we must evaluate the integral in (4.15), which is an analytically intractable problem. Thus, we used the technique proposed in [62], [60] to approximate this integral. This method produces a Monte Carlo estimate by sampling from the priors of $\vec{\alpha_j}$, $\vec{\beta_j}$ and $\vec{\gamma_j}$. According to the method proposed in [63], [65], we can simulate from the $\vec{\alpha_j}$, $\vec{\beta_j}$ and $\vec{\gamma_j}$ posterior distributions by applying the Metropolis-Hastings method. Since the conditional posterior given by (4.8) is not widely known, this method is used to avoid direct sampling of mixture parameters. To simulate the latent variable $\vec{\delta_j} = \{\vec{\theta_j}\}$ from its posterior distribution, we employ the Metropolis-Hastings approach, a widely used technique for Bayesian inference. A crucial initial step in this methodology is the definition of a suitable proposal distribution that facilitates the exploration of the posterior distribution. In the present study, we adopt a random walk Metropolis-Hastings (M-H) approach to propose new values for $\vec{\delta_j}$. Specifically, our chosen proposal distribution is expressed as follows for all positive $\tilde{\delta}_{jd} > 0$, where d ranges from $d = 1, \ldots, D$:

$$\tilde{\delta}_{jd} \sim \mathcal{LN}(\log(\delta_{jd}^{(t-1)}), \sigma^2) \tag{4.16}$$

$\mathcal{LN}(\log(\delta_{jd}^{(t-1)}), \sigma^2)$ is the log-normal distribution with mean $\log(\delta_{jd}^{(t-1)})$ and variance σ^2. Note that (4.16) can be expressed as $\log(\tilde{\delta}_{jd}) = \log(\delta_{jd}^{(t-1)}) + \epsilon_1$,

Algorithm 3 Nonparametric Bayesian learning of MBMM

1. **Process**
2. initialize assignments and parameters
3. **repeat**
4. Generate $\vec{Z}_i$ and then update n_j, for $i = 1, \ldots, N$, $j = 1, \ldots, M$.
5. Update the number of represented components denoted by M.
6. Update the mixing parameters for the represented components by $\pi_j = \frac{n_j}{N+\eta}$, $j = 1, \ldots, M$
7. Update the mixing parameters $\pi_U = \frac{\eta}{N+\eta}$ of the unrepresented clusters.
8. Generate the mixture parameters $\vec{\alpha_j}$, $\vec{\beta_j}$ and $\vec{\gamma_j}$ from (4.8) for $j = 1, \ldots, M$ using Metropolis-Hastings
9. **until** Convergence

where $\epsilon_1 \sim \mathcal{N}(0, \sigma^2)$. In the subsequent phase of the Metropolis-Hastings (M-H) algorithm, it is imperative to establish an acceptance ratio r to determine whether the newly generated samples at iteration t should be accepted or rejected for the subsequent iteration. The acceptance ratio is defined as follows:

$$k = \frac{\pi(\tilde{\delta}_j \mid \mathcal{Z}, \mathcal{X}) \prod_{d=1}^{D} \mathcal{LN}((\delta_{jd}^{(t-1)}) \mid \log(\tilde{\delta}_{jd}), \sigma^2)}{\pi(\vec{\delta}_j^{(t-1)} \mid \mathcal{Z}, \mathcal{X}) \prod_{d=1}^{D} \mathcal{LN}(\tilde{\delta}_{jd} \mid \log(\delta_{jd}^{(t-1)}), \sigma^2)} \tag{4.17}$$

In outlining the entire Bayesian learning method proposed in this study, Algorithm 3 provides a step-by-step procedure based on the Metropolis-Hastings-within-Gibbs sampler.

4.3 Experimental Results

In this part of our research, we implemented our proposed algorithm across three real-world medical images involving lung tissue samples, histological

breast tissue images, and histological malaria images. To effectively apply our models to these datasets, a preliminary step involves extracting features from images, accompanied by requisite preprocessing. Feature extraction was conducted utilizing the Scale-invariant Feature Transform (SIFT) [42,66] and Bag of Visual Words (BoVW) [43]. The Bag of Visual Words method represents images as patches, with distinctive patterns derived from each image. SIFT served as the feature detector in this process. As part of the preprocessing phase, we normalized the datasets using the min-max method ($X = \frac{X - X_{min}}{X_{max} - X_{min}}$), presuming that all input values fall within the range of (0,1). We assess the efficacy of our proposed algorithm against FLNBMM using the MCMC-based Bayesian approach, IGMM, and EM-based GMM. The accuracy is assessed through a confusion matrix comparing predicted labels with the actual ones. To evaluate our model performance against other methods, we use standard metrics based on the confusion matrix, which are defined as follows: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$. In this context, TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) denote the respective total counts, providing a comprehensive representation of the classification outcomes.

4.3.1 Malaria Detection

Malaria is potentially a fatal disease and it is essential to be diagnosed quickly and accurately [46]. One of the most common methods of diagnosing malaria is through microscopic inspection. Microscopists use a "blood smear technique". By examining the amount and shape of blood cells, they can detect the presence of malaria-infected cells. Infected cells have a small clot, while uninfected cells do not. For reference, some images of blood cells are pictured in Fig 4.2. The process of diagnosing diseases through analyzing smears is a complex as it requires the examination of a large number of samples with utmost accuracy. So, integration of a reliable machine learning approach can offer significant benefits in detecting such diseases. In our research, we evaluated the performance of our model using a cell image dataset of 2000 samples obtained from the National Institutes of Health.

4.3.2 Lung Cancer Diagnosis

Lung cancer is one of the most common causes of cancer-related deaths worldwide, and early detection is crucial for the patient's survival. Lung cancer can spread quickly if not detected in the early stages. Therefore, an accurate diagnosis is necessary to ensure timely treatment and increase the chances of a successful recovery. In our experiment's second phase, we applied our model to analyze lung histopathological images [48]. Our dataset contained 1500 images that featured benign, adenocarcinoma, and squamous cell carcinoma tissues related to lung cancer. To provide a better understanding, we have included an example of each class in Fig. 4.3.

4.3.3 Breast Tissue Analysis

Breast cancer represents a significant health concern for women worldwide. Unfortunately, it is often a leading contributor to cancer-related fatalities [44]. Early detection is crucial in improving survival rates. Despite existing strategies, current approaches to combat breast cancer have not been as practical as desired. Evaluating breast cancer histology involves analyzing the chemical and cellular features of cells from a suspicious tumor, allowing for the distinction between benign and malignant lesions based on these histological characteristics [67]. Fig. 4.4 exhibits some breast tissue samples.The provided image samples of breast tissues demonstrate that darker, more violet-colored tissue is more likely to be cancerous than those with rose colors. However, it is worth noting that this is only sometimes the case, making the learning model more challenging. The use of machine learning techniques can improve the precision of breast cancer diagnosis and reduce the number of false diagnoses. We tested our method on a publicly available dataset with 500 samples for each malignant and benign label.

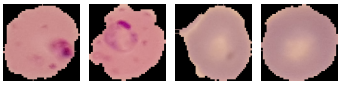

Figure 4.2: Samples of infected and uninfected cells.

Table 4.1: Results on Malaria Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
ILNBMM	89.48	88.16	91.20	89.65
FLNBMM	87.45	85.87	89.65	87.72
FGMM	85.80	84.66	87.45	86.03
IGMM	85.25	84.32	86.60	85.45

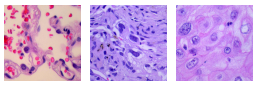

Figure 4.3: Images of three types of lung tissues.

Table 4.2: Results on Lung Cancer Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
ILNBMM	86.27	90.10	89.20	89.65
FLNBMM	84.13	89.12	86.80	87.94
FGMM	80.00	88.97	79.90	84.19
IGMM	79.47	88.53	79.50	83.77

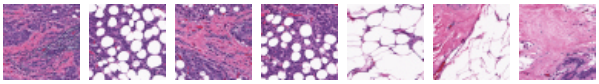

Figure 4.4: Benign and malignant samples of breast tissue.

Table 4.3: Results on Breast Tissue Dataset

Method	***Accuracy***	***Precision***	***Recall***	***F*1*-score***
ILNBMM	84.60	87.28	81.00	84.02
FLNBMM	82.60	84.24	80.20	88.17
FGMM	77.07	86.12	78.20	81.97
IGMM	76.00	85.04	77.61	81.15

Chapter 5

Conclusion

The focus of this thesis was to develop a range of unsupervised methods and apply them to medical problems with the primary objective of providing potent alternatives to commonly utilized models such as the Gaussian Mixture Model. Our proposed models are based on Libby-Novick Beta (LNB), which is a member of the Beta family. In comparison to the conventional Beta distribution, the LNB has more shape parameters, providing considerable flexibility when fitting data. In real-world data modeling, LNB can capture skewness and kurtosis of data because of this characteristic [68]. According to previous research [34, 37], LNB mixture model (LNBMM) has demonstrated convincing results in some applications including pattern recognition and image analysis.

First, we introduced a novel finite mixture model that is based on the Libby-Novick Beta distribution. To estimate the parameters of our mixture model, we utilized a deterministic methodology, maximum likelihood, via an expectation maximization algorithm and Newton-Raphson was used to update the parameters iteratively. To evaluate our model we used three medical applications and compared its performance with two other mixture models: Gaussian mixture model and Beta mixture model. Our results indicate that Libby-Novick Beta mixture model performs better than conventional and commonly used methods.

Then we have discussed a method for selecting the number of components in Libby-Novick Beta mixtures based on MML. According to the results, the

MML model selection method performs well on real-world data. The reason for this can be attributed to the fact that the prior term in this criterion is present in a manner that is not present in the other criteria. Our algorithm was evaluated for three medical applications at the end.

Finally, we proposed a nonparametric Bayesian framework for multivariate Libby-Novick Beta mixture models. In particular, we used the Dirichlet process to extend the finite model to the infinite case in order to determine the optimal number of components. The motivation for employing Bayesian learning for parameter estimation is in its ability to integrate prior knowledge about parameters, hence mitigating the risks of under- or over-fitting. In practical implementations, our proposed framework demonstrated superior performance when compared to both GMM and FLNBMM.

To conclude, our proposed models are robust and flexible according to the results of our real-world experiments.
For future research directions, incorporating feature selection into our proposed frameworks could be beneficial. Additionally, evaluating our Bayesian models across different applications could provide a more comprehensive understanding of their performance.

List of References

[1] Ahmed Hosny. Artificial intelligence in radiology. *Journal of Medical Imaging*, 15(3):45–56, 2022.

[2] Yevgeniy Kuchin and Juris Grundspeņķis. Machine learning methods for identifying composition of uranium deposits in kazakhstan. *Applied Computer Systems*, 22(1):21–27, 2017.

[3] Giuseppe Scalia. Machine learning for scientific data analysis. *Special Topics in Information Technology*, pages 115–126, 2022.

[4] Hany Alashwal, Maha E. Halaby, Jacob J. Crouse, Ahmed Abdalla, and Ahmed A. Moustafa. The application of unsupervised clustering methods to alzheimer's disease. *Frontiers in Computational Neuroscience*, 13, 2019.

[5] Nizar Bouguila. A model-based discriminative framework for sets of positive vectors classification: Application to object categorization. In *2014 1st International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 277–282, 2014.

[6] Ziyang Song, Samr Ali, and Nizar Bouguila. Background subtraction using infinite asymmetric gaussian mixture models with simultaneous feature selection. *IET Image Process.*, 14(11):2321–2332, 2020.

[7] Ali Sefidpour and Nizar Bouguila. Spatial color image segmentation based on finite non-gaussian mixture models. *Expert Syst. Appl.*, 39(10):8993–9001, 2012.

[8] John Smith and Mary Johnson. Recent advances in finite mixture models for data analysis. *Journal of Statistical Learning*, 8(2):123–145, 2022.

[9] Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

[10] Wentao Fan, Faisal R. Al-Osaimi, Nizar Bouguila, and Ji-Xiang Du. Proportional data modeling via entropy-based variational bayes learning of mixture models. *Appl. Intell.*, 47(2):473–487, 2017.

[11] Lin Yang, Wentao Fan, and Nizar Bouguila. Clustering analysis via deep generative models with mixture models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1):340–350, 2022.

[12] Nizar Bouguila, Khaled Almakadmeh, and Sabri Boutemedjet. A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection. *Expert Syst. Appl.*, 39(7):6641–6656, 2012.

[13] Nizar Bouguila and Wentao Fan. *Mixture models and applications*. Springer, 2020.

[14] H. Huang, Z. Liao, X. Wei, and Y. Zhou. Combined gaussian mixture model and pathfinder algorithm for data clustering. *Entropy*, 25:946, 2023.

[15] Shuai Fu and Nizar Bouguila. Asymmetric gaussian mixtures with reversible jump mcmc. In *2018 IEEE Canadian Conference on Electrical Computer Engineering (CCECE)*, pages 1–4, 2018.

[16] Can Hu, Wentao Fan, Ji-Xiang Du, and Nizar Bouguila. A novel statistical approach for clustering positive data based on finite inverted beta-liouville mixture models. *Neurocomputing*, 333:110–123, 2019.

[17] Mahsa Amirkhani, Narges Manouchehri, and Nizar Bouguila. Birth-death mcmc approach for multivariate beta mixture models in medical applications. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 285–296. Springer, 2021.

[18] Narges Manouchehri, Nizar Bouguila, and Wentao Fan. Batch and online variational learning of hierarchical pitman-yor mixtures of multivariate beta distributions. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 298–303. IEEE, 2021.

[19] Narges Manouchehri and Nizar Bouguila. Stochastic expectation propagation learning of infinite multivariate beta mixture models for human tissue analysis. In *IECON 2021–47th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6. IEEE, 2021.

[20] Narges Manouchehri, Nizar Bouguila, and Wentao Fan. Nonparametric variational learning of multivariate beta mixture models in medical applications. *International Journal of Imaging Systems and Technology*, 31(1):128–140, 2021.

[21] Nizar Bouguila, Djemel Ziou, and Riad I. Hammoud. A bayesian non-gaussian mixture analysis: Application to eye modeling. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[22] Yuan Ji, Chunlei Wu, Ping Liu, Jing Wang, and Kevin R Coombes. Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122, 2005.

[23] Nizar Bouguila and Djemel Ziou. Dirichlet-based probability model applied to human skin detection [image skin detection]. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–521. IEEE, 2004.

[24] Rua Alsuroji, Nuha Zamzami, and Nizar Bouguila. Model selection and estimation of a finite shifted-scaled dirichlet mixture model. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 707–713, 2018.

[25] Wentao Fan, Hassen Sallay, and Nizar Bouguila. Online learning of hierarchical pitman–yor process mixture of generalized dirichlet distributions with feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 28(9):2048–2061, 2017.

[26] Bromensele Samuel Oboh and Nizar Bouguila. Unsupervised learning of finite mixtures using scaled dirichlet distribution and its application to software modules categorization. In *2017 IEEE International Conference on Industrial Technology (ICIT)*, pages 1085–1090, 2017.

[27] Nizar Bouguila and Djemel Ziou. Using unsupervised learning of a finite dirichlet mixture model to improve pattern recognition applications. *Pattern Recognit. Lett.*, 26(12):1916–1925, 2005.

[28] Darya Forouzanfar, Narges Manouchehri, and Nizar Bouguila. A fully bayesian inference approach for multivariate mcdonald's beta mixture model with feature selection. In *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 2055–2060, 2023.

[29] Tong Zhang. Statistical machine learning: A comprehensive review. *Statistica Sinica*, 28(4):1595–1636, 2018.

[30] Teresa Cousineau, Denis Allan. Likelihood and its use in parameter estimation and model comparison. *Mesure et évaluation en éducation*, 37(3):63–98, 2015.

[31] Shuai Fu and Nizar Bouguila. A bayesian intrusion detection framework. In *2018 International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2018, Glasgow, Scotland, United Kingdom, June 11-12, 2018*, pages 1–8. IEEE, 2018.

[32] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[33] Tarek Elguebaly and Nizar Bouguila. Simultaneous bayesian clustering and feature selection using rjmcmc-based learning of finite generalized dirichlet mixture models. *Signal Process.*, 93(6):1531–1546, 2013.

[34] Niloufar Samiee, Narges Manouchehri, and Nizar Bouguila. Maximum likelihood-based estimation of finite multivariate libby-novick beta mixture models in medical applications. In *IEEE International Conference on Industrial Technology, ICIT 2023, Orlando, FL, USA, April 4-6, 2023*, pages 1–7. IEEE, 2023.

[35] Niloufar Samiee, Narges Manouchehri, and Nizar Bouguila. Finite libby-novick beta mixture model: An mml-based approach. In Ngoc Thanh Nguyen, Siridech Boonsang, Hamido Fujita, Bogumila Hnatkowska, Tzung-Pei Hong, Kitsuchart Pasupa, and Ali Selamat, editors, *Intelligent Information and Database Systems - 15th Asian Conference, ACIIDS 2023, Phuket, Thailand, July 24-26, 2023, Proceedings, Part I*, volume 13995 of *Lecture Notes in Computer Science*, pages 371–383. Springer, 2023.

[36] Niloufar Samiee, Narges Manouchehri, and Nizar Bouguila. A nonparametric bayesian framework for multivariate libby-novick beta mixture models. In *10th International Conference on Control, Decision and Information Technologies, CoDIT 2024, Valetta, Malta, July 1-4, 2024*, 2024. submitted.

[37] Kian Ketabchi, Narges Manouchehri, and Nizar Bouguila. Fully bayesian libby-novick beta mixture model with feature selection. In *IEEE International Conference on Industrial Technology, ICIT 2022, Shanghai, China, August 22-25, 2022*, pages 1–6. IEEE, 2022.

[38] Gauss Cordeiro, Luis Santana, Edwin Ortega, and Rodrigo Pescim. A new family of distributions: Libby-novick beta. *International Journal of Statistics and Probability*, pages 63–80, 2014.

[39] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[40] Robert I Jennrich and Stephen M Robinson. A newton-raphson algorithm for maximum likelihood factor analysis. *Psychometrika*, 34:111–123, 1969.

[41] Jai Puneet Singh and Nizar Bouguila. Proportional data clustering using k-means algorithm: A comparison of different distances. In *2017 IEEE International Conference on Industrial Technology (ICIT)*, pages 1048–1052, 2017.

[42] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[43] L. Jagjeevan Rao, P. Neelakanteswar, Madupu Ramkumar, Azmira Krishna, and CMAK Zeelan Basha. An effective bone fracture detection using bag-of-visual-words with the features extracted from sift. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 6–10, 2020.

[44] Angela N Giaquinto, Hyuna Sung, Kimberly D Miller, Joan L Kramer, Lisa A Newman, Adair Minihan, Ahmedin Jemal, and Rebecca L Siegel. Breast cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(6):524–541, 2022.

[45] https://www.kaggle.com/paultimothymooney/breast-histopathology-images.

[46] Noppadon Tangpukdee, Chatnapa Duangdee, Polrat Wilairatana, and Srivicha Krudsood. Malaria diagnosis: A brief review. *The Korean journal of parasitology*, 47:93–102, 07 2009.

[47] *Malaria dataset*. https://ceb.nlm.nih.gov/repositories/malaria-datasets.

[48] *Lung dataset*, 2018. https://www.kaggle.com/andrewmvd/lung-and-colon-cancer-histopathological-images.

[49] N. Bouguila and D. Ziou. Mml-based approach for high-dimensional unsupervised learning using the generalized dirichlet mixture. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 53–53, 2005.

[50] Bengt Autzen. Bayesian ockham's razor and nested models. *Economics and Philosophy*, 35(2):321–338, 2019.

[51] https://www.cancer.gov/types/breast.

[52] https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/risk-factors.html.

[53] Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün. Model-based clustering based on sparse finite gaussian mixtures. *Statistics and computing*, 26(1-2):303–324, 2016.

[54] Nizar Bouguila, Djemel Ziou, and Ernest Monga. Practical bayesian estimation of a finite beta mixture through gibbs sampling and its applications. *Statistics and Computing*, 16(2):215–225, 2006.

[55] Wentao Fan and Nizar Bouguila. Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE Trans. Neural Networks Learn. Syst.*, 24(11):1850–1862, 2013.

[56] Tarek Elguebaly and Nizar Bouguila. Bayesian learning of finite generalized gaussian mixture models on images. *Signal Process.*, 91(4):801–820, 2011.

[57] Tarek Elguebaly and Nizar Bouguila. Bayesian learning of generalized gaussian mixture models on biomedical images. In Friedhelm Schwenker and Neamat El Gayar, editors, *Artificial Neural Networks in Pattern Recognition, 4th IAPR TC3 Workshop, ANNPR 2010, Cairo, Egypt, April 11-13, 2010. Proceedings*, volume 5998 of *Lecture Notes in Computer Science*, pages 207–218. Springer, 2010.

[58] Wentao Fan and Nizar Bouguila. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In Francesca Rossi, editor, *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 1323–1329. IJCAI/AAAI, 2013.

[59] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 2013.

[60] Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

[61] Nizar Bouguila and Djemel Ziou. A dirichlet process mixture of dirichlet distributions for classification and prediction. In *2008 IEEE workshop on machine learning for signal processing*, pages 297–302. IEEE, 2008.

[62] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[63] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[64] Nizar Bouguila. Infinite liouville mixture models with application to text and texture categorization. *Pattern Recognition Letters*, 33(2):103–110, 2012.

[65] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[66] Philipp Kreowsky and Benno Stabernack. A full-featured fpga-based pipelined architecture for sift extraction. *IEEE Access*, 9:128564–128573, 2021.

[67] S. Wang, Y. Wei, Z. Li, J. Xu, and Y. Zhou. Development and validation of an mri radiomics-based signature to predict histological grade in patients with invasive breast cancer. *Breast Cancer (Dove Medical Press)*, 14:335–342, 2022.

[68] Zafar Iqbal, Muhammad Rashad, and Muhammad Hanif. Properties of the libby-novick beta distribution with application. *International Journal of Analysis and Applications*, 19:360–388, 04 2021.